EDGE-BASED DEPTH GRADIENT REFINEMENT FOR 2D TO 3D LEARNED PRIOR CONVERSION

José L. Herrera *Carlos R. del-Blanco* *Narciso García*

ABSTRACT

2D-to-3D conversion is an important task for reducing the current gap between the number of 3D displays and the available 3D content. Here, we present an automatic 2D-to-3D image conversion approach based on machine learning principles. Stemming from the hypothesis that images with a similar structure have likely a similar 3D structure, the depth of a query color image is estimated using a color plus depth image dataset. Clusters with common scene structure are computed offline. Then, a matching process is performed to select the cluster centroid which is the most similar to the query image. A prior depth map is computed fusing the depth maps of the images in this cluster. Then, an edge-based post-processing stage is applied to the prior depth map estimation to enhance the final scene depth estimation. Promising results are obtained in two commonly used databases achieving a similar performance to other much complex state-of-the-art approaches.

Index Terms — 2D-to-3D conversion, depth maps, depth prior, clustering, machine learning

1. INTRODUCTION

An important rise in the number of 3D players and displays such as TVs, smartphones, cinemas, DVD/Blu-Ray or video game consoles has happened in the last decade. Despite of this, the amount of available 3D content, like images, movies or TV broadcasting, has not increased at the same rate creating an important gap between the quantity of 3D players and the volume of 3D content. To balance this situation, different algorithms have appeared to convert, automatically or semi-automatically, the current 2D content into 3D to satisfy the demand of the users for 3D experience.

The 2D-to-3D image and video conversion is a task that is usually performed in two main stages. First, a depth map of the image or video is estimated, and then, this estimation and the original image are used to render a stereoscopic pair. Here, we focus in the first stage of the algorithm: depth extraction from a monocular image, which is more challenging, and in addition, there exists many algorithms that generate a stereo-pair that achieve a good quality.

During the last years, new machine learning-based algorithms have appeared as an alternative for the 2D to 3D image and video conversion task. The usual hypothesis behind these new methods is that images which have a high photometrical similarity will probably have similar 3D structures (depths). Saxena et al [1] performed a supervised learning strategy to estimate the scene structure from a monocular image using an image parsing strategy and Markov Random Fields to determine 3D locations and orientations. In [2][3], the incorporation of semantic labels and more sophisticated methods achieved better scene depth results. In [4], Karsch et al. used an approach based on Scale Invariant Feature Transform (SIFT) flow and an optimization post-process to improve the results and extend the method to work with videos. Konrad et al. [5] proposed a more computationally efficient method using a descriptor based on Histogram of Oriented Gradients (HOG) to match similar images instead of the SIFT flow approach. Also a Joint Bilateral Filter is used to enhance the resulting depth map. In our previous work [6], a new approach based on Local Binary Patterns (LBP) features are used to find an adaptive number of similar images that are fused in a weighted way to estimate the depth scene structure. Since the computational cost of these methods is proportional to the size of the database, these algorithms become impractical when using large databases. To alleviate this problem, in [7] we also presented an approach based on hierarchical search in clustered databases, which improves the efficiency of the search process. In [8] we used GIST as a descriptor since it improves the performance of the algorithm, and we modified the search for finding similar images by means of a saliency-based weighting strategy to get better results in the regions of the scene with higher visual demand. Other strategies, such as the presented in [9], avoid the search for similar images by using directly a depth prior selected from a built-in depth priors set, and using a segmentation-based filtering to enhance the scene depth structure.

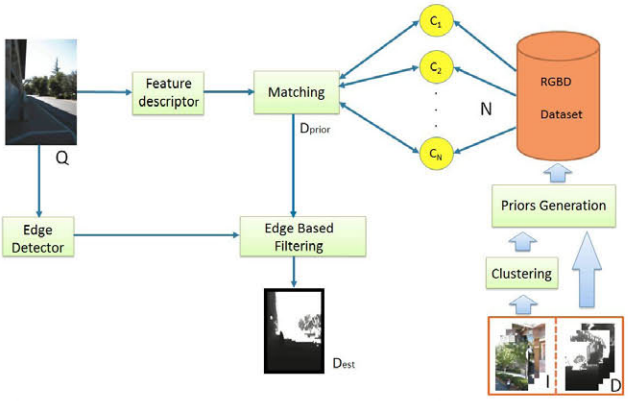

Figure 1. Block diagram of the proposed 2D-to-3D image conversion algorithm

Those methods previously presented usually lack of well defined edges. To solve these problems, while keeping a reasonable computational complexity and fast search times, a new machine learning and automatic 2D-to-3D image conversion algorithm is proposed. The conversion of a query image is performed in three main parts. First of all, before the conversion process starts, a database of pairs of images and depth maps is divided into clus-

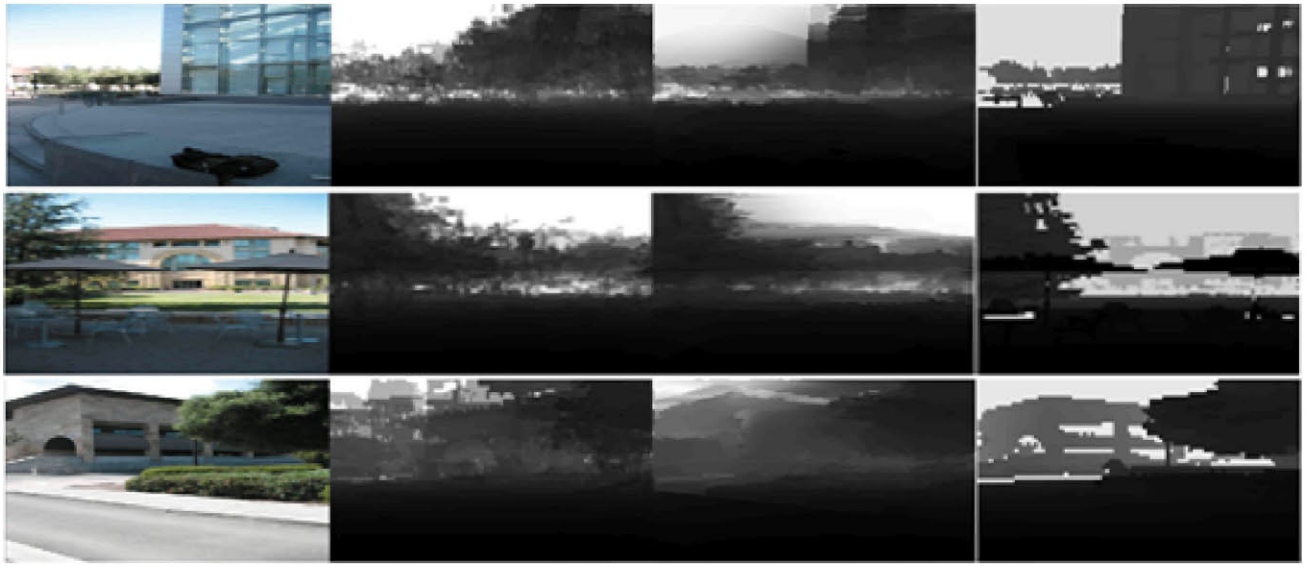

Figure 2. From left to right: Query Image, Depth Map Prior, Final Depth Map Estimation and Depth Ground Truth for images of Make3D dataset

ters using the GIST [10] descriptor, which provides a global representation of the scene instead of local details as the previously referred descriptor do. Then, a depth map prior for the query image is estimated using a machine learning approach. The algorithm matches, using the GIST descriptor, the query image with the cluster centroids estimated in the offline process, and gets from the matched cluster a depth prior which is the result of the combination of the depth maps associated to the images in that clusters by the application of the median operator. In the last step, an edge-based refinement post-processing is performed by applying a size adaptive filter to enhance the 3D structure of the scene.

2. ALGORITHM DESCRIPTION

Given a color query image Q, and a database DB composed by pairs of color images I and their associated depth maps D, the purpose of the presented approach is the estimation of depth map D_{est} of Q. The correct performance of the algorithm is subject to the presence of structurally similar images to Q in DB.

The proposed 2D-to-3D image conversion algorithm can be divided into three main stages. In the first stage, that can be considered a pre-processing stage, a clustering of all color images in the database is performed to make clusters containing similar images from a photometrical point of view and assign a depth map prior to each cluster. In the second stage, the query image is matched with the centroids of all the clusters computed in the previous stage and the depth prior corresponding to the cluster that matches best the query image is selected. In the third stage, an heuristic edge-based post-processing is applied to the selected depth prior in order to enhance the 3D structure of the scene. The block diagram of the proposed system can be shown in Fig. 1.

The difference with our previous approaches [6][7] is the learning based prior generations performed during the clustering and the edge-based post-processing stage that enhances the scene depth map estimation.

2.1. Database Clustering

The clustering process has a double goal. First, the images in the database are organized by structural similarity. In parallel, dividing the database into clusters lets the algorithm work in a more efficient way in large databases as it was described in [7]. This stage of the algorithm can be performed offline, before a query image Q arrives for conversion.

The clustering is performed over a compact feature-based representation of the color images in the dataset. This feature descriptors are based on GIST [10] descriptor and are built by dividing the image into 4 x 4 tiles, and obtaining the GIST descriptor in each of these tiles. Then, the descriptors of each tile are concatenated in a single vector F_I which characterizes the whole image. GIST has been selected to be used for building the descriptor since it provides a global description of the scene.

The k-means algorithm along with the correlation coefficient, as similarity metric, is used to cluster the color images according to their feature-based representation in such a way that images grouped in the same cluster have a similar structure. The centroid of each cluster F_C is the average across all the GIST-based feature descriptors of the color images in the cluster

Once the clusters have been obtained, a depth map prior is assigned to each cluster by computing the median operator across the depth maps associated to all images grouped in the cluster.

2.2. Matching

The matching process starts with the computation of the GIST descriptor of the query image F_Q. This descriptor computation is performed in the same way that is done for the images in the dataset, but in this case is an online process that cannot be performed beforehand. Then, the similarity between the descriptor of the query image and the centroids of the clusters of the database DB is computed using correlation as the similarity metric as follows:

$$c(n) = corr(F_Q, F_{Cn}), \tag{1}$$

where $corr()$ is the correlation measure, F_Q is the GIST-based image feature descriptor of the query image Q, and F_{Cn} is the n^{th} centroid of the clustered database DB.

The depth map prior, associated to the cluster with the highest similarity is chosen as the depth map prior D_{prior} of the query image Q.

2.3. Edge-based Post-processing

The depth map prior previously estimated works as a first approximation of the depth map of the scene, following the general vari-

Figure 3. From left to right: Query Image, Depth Map Prior, Final Depth Map Estimation and Depth Ground Truth for images of NYU Kinect dataset

ations of the depth of the image but it needs to be enhanced to obtain a good depth estimation of the images where the depth of the different objects could be appreciated and not just global variations.

The aim of this stage of the algorithm is to delimitate the depth of the different objects of the color image and to smooth the depth variations inside the different objects in the scene

Based on the hypothesis that edges in a color image and in a depth map use to match, we perform over the depth prior an heuristic edge-based post-processing to refine the depth map prior and enhance the 3D structure of the scene. To perform this post-processing, we first compute the Sobel edge detector of the query image Q.

Then, we apply an adaptive size filter with a Gaussian kernel. The size of this filter changes to cover always a large area of the image but without reaching any of the pixels detected as an edge by the Sobel detector. The initial window shape for pixel p is the largest square which do not include any of the pixels detected as edges by Sobel. Then, the size of the window grows in each direction until reaches the pixels detected as edges. A Gaussian kernel is applied to the pixels in the resulting window centered in pixel p.

3. EXPERIMENTAL RESULTS

The proposed approach has been tested in two different datasets. The Make3D dataset [11], composed by 534 pairs of color images and their corresponding depths, and the NYU Kinect database [12], with a total of 1449 pairs of images + depth maps.

The resolution of the Make3D dataset is 2272 x 1704 for color images and 55 x 305 for depth maps, while in NYU database, the resolutions is 640 x 480 for both color images and depth maps. For a straightforward comparison with other approaches in the state of the art, images and depth maps of both databases have been resized to 320 x 240 pixels.

The quality of the final estimation is sensible to the value of the number of clusters in which the databases are divided. We have chosen a value of 75 clusters for the Make3D dataset and 80 clusters for the NYU database, as these values achieve the best performance.

To evaluate the performance of the algorithm quantitatively, we performed the tests in a leave-one-out cross-validation configuration for both databases. As the quality metric, we used the correlation coefficient (C), the Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM), all of them computed between the depth ground truth and the depth estimation provided by this method. The correlation coefficient is defined as follows:

$$C = \frac{\sum_i (D_{est}[i] - \mu_{D\,est})(D_Q[i] - \mu_{D\,Q})}{N\sigma_{D_{est}}\sigma_{D_Q}}, \tag{2}$$

where N is the number of pixels in $\mu_{D\,est}$ and D_Q (ground-truth depth of the query image Q), $\mu_{D\,est}$ and $\mu_{D\,Q}$ are the empirical mean values of D_{est} and D_Q, respectively, $\sigma_{D_{est}}$ and σ_{d_Q} are the corresponding empirical standard deviations, and it refers to each pixel of the image. The normalized cross-covariance C takes values from -1 to +1 (values close to +1 indicate that the depth maps are very similar an values close to -1 suggest they are complementary). The other metricsare mathematically expressed by

$$\begin{aligned} RMSE &= \sqrt{\sum_i (D_Q(i) - D_{est}(i))^2 / N}, \\ PSNR &= 20 \log_{10} \frac{max(D_Q)}{RMSE}, \end{aligned} \tag{3}$$

where max is a function that return the maximum value.

The SSIM is used as proposed in [13]. The higher the values of these metrics are, the higher quality has been achieved in the estimation

The results of our approach have been compared with the Depth Transfer approach of Karsch et al. [4] and the HOG-Based Depth Learning approach of Konrad et al [5]. Tables 1 and 2 show the quantitative results for Make3D [11] and NYU [12] databases. As can be seen, for Make3D dataset, we achieve a similar quality than the other algorithms (best for PSNR and second best for the other C and SSIM) while for NYU dataset, we outperform the other two approaches for the three used metrics.

In Fig. 2 and 3, some examples of query images, depth map priors and final depth maps estimations, generated by this method, are shown.

While NYU dataset is composed by indoor scenes, Make3D images represent outdoor scenes, which tend to have smoother

Make3D [11]			
Algorithm — metric	C	PSNR	SSIM
Depth Transfer [4] (2012)	0.66	13.8	0.82
HOG Learning Based [5] (2012)	0.58	14.0	0.79
Adaptive LBP-based [6] (2014)	0.63	14.4	0.77
Edge based refinement (Ours)	0.64	14.5	0.80

Table 1. Evaluation of state-of-the-art algorithms using the Correlation Coefficient (C), PSNR and Structural Similarity (SSIM) metrics in the Make3D database in Leave One Out configuration. The results are the average over the 534 test images.

NYU [12]			
Algorithm - metric	C	PSNR	SSIM
Depth Transfer [4] (2012)	0.60	12.0	0.78
HOG Learning Based [5] (2012)	0.56	12.9	0.79
Adaptive LBP-based [6] (2014)	0.61	13.5	0.80
Edge based refinement (Ours)	0.63	13.7	0.81

Table 2. Evaluation of state-of-the-art algorithms using the Correlation Coefficient (C), PSNR and Structural Similarity (SSIM) metrics in the NYU database in Leave One Out configuration. The results are the average over the 1449 test images.

depth maps, easier to estimate even with a smaller database. As it is showed in Tables 1 and 2, the decrease in the quality measures of our approach is less pronounced than the reduction suffered by the other algorithms. Especially significant is the reduction in the performance of Karsch approach[4] for the indoor images dataset.

Therefore, we outperform the HOG-Based Depth Learning approach (C, PSNR, and SSIM) and the Depth Transfer approach (PSNR) and we have slightly better results than Depth Transfer approach (C and SSIM), but a significantly lower computational cost.

4. CONCLUSIONS

A novel algorithm for automatically estimate the 3D structure of a query image has been presented in this paper. The approach, uses a machine learning framework to infer the depth of a given image using a database composed by pairs of color images and depth maps. Our method uses K-means to divide the database into different clusters and assign to each cluster a depth map prior based on the depth maps of the images in the cluster. This clustering also allows to extend the use of the algorithm to larger databases avoiding the problem of excessively long times that present this kind of algorithms for the search stage. The query image is matched with all the cluster centroids and the depth map prior of the cluster that matches best with the query image is selected as a depth map prior of the query image. An edge-based refinement post-process is applied to the depth map prior to enhance the 3D structure of the scene and obtain a good final depth estimation for the query image. The algorithm keeps a constant quality value for the two used dataset, getting similar results than others state-of-art approaches in the outdoor database and outperforms them with the indoor database, whose depths are harder to estimate, while keeping a reduced complexity.

5. REFERENCES

[1] A. Saxena, H. Chung Sung, and Y. Ng Andrew, "Learning depth from single monocular images," in *NIPS 18*. 2005, MIT Press.

[2] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *IEEE Conf. on Comput. Vis. and Pattern Recognit., CVPR 2009.*, June 2009, pp. 1972–1979.

[3] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *IEEE Conf. on Comput. Vis. and Pattern Recognit., CVPR 2010.*, June 2010, pp. 1253–1260.

[4] K. Karsch, C. Liu, and S. Kang, "Depth extraction from video using non-parametric sampling," in *Computer Vision ECCV 2012*, 2012, vol. 7576 of *Lecture Notes in Computer Science*, pp. 775–788.

[5] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning-based, automatic 2D-to-3D image and video conversion," *IEEE Trans. on Image Process.*, vol. 22, no. 9, pp. 3485–3496, Sept 2013.

[6] J.L. Herrera, C.R. del Blanco, and N. Garcia, "Learning 3D structure from 2D images using LBP features," in *IEEE International Conference on Image Processing*, October 2014, pp. 2022–2025.

[7] J.L. Herrera, C.R. del Blanco, and N. Garcia, "Fast 2d to 3d conversion using a clustering-based hierarchical search in a machine learning framework," in *IEEE 3DTV-Conference*, July 2014, pp. 1–4.

[8] J.L. Herrera, J. Konrad, C.R. del Blanco, and N. Garcia, "Learning-based depth estimation from 2D images using GIST and saliency," in *IEEE International Conference on Image Processing*, September 2015.

[9] C. Li C. Cheng and L. Chen, "A novel 2D-to-3D conversion system using edge information," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1739–1745, Aug 2010.

[10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[11] A. Saxena, M. Sun, and A.Y. Ng, "Make3d: Learning 3D scene structure from a single still image," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

[12] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011*, Nov 2011, pp. 601–608.

[13] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.